



LAW OFFICES  
SUGHRUE, MION, ZINN, MACPEAK & SEAS, PLLC  
1010 EL CAMINO REAL  
MENLO PARK, CA 94025  
TELEPHONE (650) 325-5800  
FACSIMILE (650) 325-6606



WASHINGTON DC OFFICE  
2100 PENNSYLVANIA AVENUE, N.W.  
WASHINGTON, D.C. 20037-3203  
TELEPHONE (202) 293-7060  
FACSIMILE (202) 293-7860

JAPAN OFFICE  
TOEI NISHI SHIMBASHI BLDG. 4F  
13-5 NISHI SHIMBASHI 1-CHOME  
MINATO-KU, TOKYO 105, JAPAN  
TELEPHONE (03) 3503-3760  
FACSIMILE (03) 3503-3756

April 24, 2000

**BOX PATENT APPLICATION**  
Assistant Commissioner for Patents  
Washington, D.C. 20231

Express Mail No. EK165357562US

Re: Application of Yihong GONG and Xin LIU  
**METHOD AND SYSTEM FOR SEGMENTATION,  
CLASSIFICATION, AND SUMMARIZATION OF VIDEO IMAGES**  
Our Ref: CA1055

Dear Sir:

Attached hereto is the application identified above including thirty-seven (37) sheets of the specification, claims, five (5) sheets of formal drawings, **executed Assignment, PTO 1595 form, and executed Declaration and Power of Attorney**, and a Preliminary Amendment.

The Government filing fee is calculated as follows:

Total claims	<u>72</u> - 20 = <u>52</u>	x	\$18 =	\$ 936.00
Independent claims	<u>12</u> - 3 = <u>9</u>	x	\$78 =	\$ 702.00
Base Fee				<u>\$ 690.00</u>
<b>TOTAL FILING FEE</b>				<b>\$ 2,328.00</b>
Recordation of Assignment				<u>\$ 40.00</u>
<b>TOTAL FEE</b>				<b>\$ 2,368.00</b>

Checks for the statutory filing fee of \$2,328.00 and Assignment recordation fee of \$40.00 are attached. You are also directed and authorized to charge or credit any difference or overpayment to Deposit Account No. 19-4880. The Commissioner is hereby authorized to charge any fees under 37 C.F.R. §§ 1.16 and 1.17 and any petitions for extension of time under 37 C.F.R. § 1.136 which may be required during the entire pendency of the application to Deposit Account No. 19-4880. A duplicate copy of this transmittal letter is attached.

SUGHRUE, MION, ZINN, MACPEAK & SEAS, PLLC

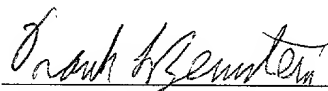
Assistant Commissioner for Patents

Page 2

Priority is claimed from November 24, 1999, and December 17, 1999, based on U.S. (Provisional) Application Nos. 60/167,230 and 60/172,379.

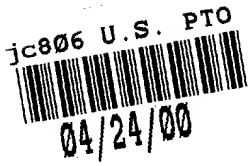
Respectfully submitted,

SUGHRUE, MION, ZINN,  
MACPEAK & SEAS, PLLC  
Attorneys for Applicant

By:   
Frank L. Bernstein  
Registration No. 31,484

Customer No. 23493

004340" 64E95603



04-25-00

A

**PATENT APPLICATION**

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re application of

Yihong GONG and Xin LIU

Filed: April 24, 2000

For: METHOD AND SYSTEM FOR SEGMENTATION, CLASSIFICATION, AND  
SUMMARIZATION OF VIDEO IMAGES

**"EXPRESS MAIL" CERTIFICATE OF MAILING**

I hereby certify that this paper and the attachments hereto are being deposited with the United States Postal Service "Express Mail Post Office to Addressee" service under 37 CFR 1.10 on the date indicated below, addressed to the Assistant Commissioner for Patents, Washington, DC 20231.

**EK165357562US**

Express Mail Certificate Number: EK165357562US

Date: April 24, 2000

Signed: Thea K. Wagner  
Thea K. Wagner

Enclosures:

1. PrintEFS bibliographic data sheet output;
2. PTO Transmittal Letter (original and 1 copy) with check No. 150892 (\$2,328.00) and the executed Declaration;
3. Application, including the specification (24 pages), claims 1-72 (pages 24-36), and the Abstract (page 37);
4. Formal Drawings (Figs. 1-5);
5. Preliminary Amendment; and
6. PTO Form 1595 with the executed Assignment and check No. 150893 (\$40.00).

INVENTOR INFORMATION

Inventor One Given Name:: Yihong  
Family Name:: GONG  
Postal Address Line One:: 503 Alberta Avenue  
City:: Sunnyvale  
State or Province:: CA  
Country:: USA  
Postal or Zip Code:: 94087  
City of Residence:: Sunnyvale  
State or Province of Residence:: CA  
Country of Residence:: USA  
Citizenship Country:: China  
Inventor Two Given Name:: Xin  
Family Name:: LIU  
Postal Address Line One:: 340 Auburn Way, #24  
City:: San Jose  
State or Province:: CA  
Country:: USA  
Postal or Zip Code:: 95129  
City of Residence:: San Jose  
State or Province of Residence:: CA  
Country of Residence:: USA  
Citizenship Country:: China

JC520 U.S. PTO  
09/556349  
04/24/00

CORRESPONDENCE INFORMATION

Correspondence Customer Number:: 23493  
Fax One:: 650-325-6606  
Electronic Mail One:: fbernstein@sughrue.com

APPLICATION INFORMATION

Title Line One:: METHOD AND SYSTEM FOR SEGMENTATION, CLAS  
Title Line Two:: SIFICATION, AND SUMMARIZATION OF VIDEO I  
Title Line Three:: MAGES  
Total Drawing Sheets:: 5  
Formal Drawings?: Yes  
Application Type:: Utility  
Docket Number:: CA1055  
Secrecy Order in Parent Appl.?: No

REPRESENTATIVE INFORMATION

Representative Customer Number:: 23493  
Registration Number One:: 31484

CONTINUITY INFORMATION



**PATENT APPLICATION**

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re the Application of

Express Mail No. EK165357562US

Yihong GONG et al.

Filed: Concurrently herewith

For: METHOD AND SYSTEM FOR SEGMENTATION, CLASSIFICATION, AND  
SUMMARIZATION OF VIDEO IMAGES

**PRELIMINARY AMENDMENT**

Assistant Commissioner of Patents  
Washington, DC 20231

Sir:

Prior to initial examination on the merits, please amend the above-identified application  
as follows.

**IN THE SPECIFICATION:**

Page 24, the last four lines, delete in their entirety.

**IN THE CLAIMS:**

Page 25, before line 1, insert the following four lines:

**--WHAT IS CLAIMED IS:**

- 1 1. A method for summarizing a content of an input video sequence, said input video
- 2 sequence comprising a plurality of frames, said plurality of frames being grouped into a plurality
- 3 of video segments, said method comprising:--

PATENT APPLICATION

PRELIMINARY AMENDMENT

REMARKS

The foregoing amendment has been made to correct a pagination error. Early, favorable consideration on the merits is respectfully requested.

Respectfully submitted,



Frank L. Bernstein

Registration No. 31,484

SUGHRUE, MION, ZINN,  
MACPEAK & SEAS, PLLC  
Tel: (650) 325-5800

Customer No. 23493

Date: April 24, 2000

# METHOD AND SYSTEM FOR SEGMENTATION, CLASSIFICATION, AND SUMMARIZATION OF VIDEO IMAGES

## BACKGROUND OF THE INVENTION

The present application claims benefit from Provisional Application No. 60/167,230, filed November 24, 1999, and Provisional Application No. 60/172,379, filed December 17, 1999. The disclosures of these provisional applications are incorporated herein by reference.

### 1. Field of the Invention

This invention relates to techniques for video summarization based on the singular value decomposition (SVD) technique. The present invention also relates to providing tools for effective searching and retrieval of video sequences according to user-specified queries. In particular, the invention relates to segmentation of video sequences into shots for automated searching, indexing, and access. Finally, this invention relates to a method for extracting of features and metadata from video shots to enable classification, search, and retrieval of the video shots.

### 2. Description of the Related Art

The widespread distribution of video information in computer systems and networks has presented both excitement and challenge. Video is exciting because it conveys real-world scenes most vividly and faithfully. On the other hand, handling video is challenging because video images are represented by voluminous, redundant, and unstructured data streams which span the time sequence. In many instances, it can be a painful task to locate either the appropriate video sequence or the desired portion of the video information from a large video data collection. The situation becomes even worse on the Internet. To date, increasing numbers of websites offer video images for news broadcasting, entertainment, or product promotion. However, with very limited network bandwidth available to most home users, people spend



minutes or tens of minutes downloading voluminous video images, only to find them irrelevant.

Important aspects of managing a large video data collection are providing a user with a quick summary of the content of video footage and enabling the user to quickly browse through extensive video resources. Accordingly, to turn unstructured, voluminous video images into exciting, valuable information resources, browsing and summarization tools that would allow the user to quickly get an idea of the overall content of video footage become indispensable.

Currently, most video browsing tools use a set of keyframes to provide content summary of a video sequence. Many systems use a constant number of keyframes for each detected scene shot, while others assign more keyframes to scene shots with more changes. There are also systems that remove redundancies among keyframes by clustering the keyframes based on their visual similarity. An important missing component in existing video browsing and summarization tools is a mechanism to estimate how many keyframes would be sufficient to provide a good, nonredundant representation of a video sequence.

Simple methods that assign a fixed number of keyframes to each scene shot suffer from poor video content representation, while more sophisticated approaches that adaptively assign keyframes according to the activity levels often rely on the user to provide either the number of keyframes to be generated, or some threshold values (e.g., the similarity distance or the time interval between keyframes), which are used to generate the keyframes. Accordingly, the user must go through several rounds of interactions with the system to obtain an appropriate set of keyframes. This approach is acceptable when the user browses a small set of video images disposed on a local workstation. On the other hand, the approach becomes prohibitive when video images located on the Internet are accessed through a network with very limited bandwidth, or when a video summary must be created for each video image in a large-scale video database.

As mentioned above, existing video browsing and content overview tools utilize keyframes extracted from original video sequences. Many works concentrate

on breaking video into shots, and then finding a fixed number of keyframes for each detected shot. For example, Tonomura et al. used the first frame from each shot as a keyframe, *see* Y. Tonomura, A. Akutsu, K. Otsuji, and T. Sadakata, "Videomap and videospaceicon: Tools for anatomizing video content," in *Proc. ACM INTERCHI'93, 1993*. Ueda et al. represented each shot by using its first and last frames, *see* H. Ueda, T. Miyatake, and S. Yoshizawa, "Impact: An interactive natural-motion-picture dedicated multimedia authoring system," in *Proc. ACM SIGCHI'91*, (New Orleans), Apr. 1991. Ferman and Tekalp clustered the frames in each shot, and selected the frame closest to the center of the largest cluster as the keyframe, *see* A. Ferman and A. Tekalp, "Multiscale content extraction and representation for video indexing," in *Proc. SPIE 3229 on Multimedia Storage and Archiving Systems II*, 1997.

An obvious disadvantage of the above equal-density-keyframe assignment is that long shots, which involve camera pans and zooms as well as the object motion, will not be adequately represented. To address this problem, DeMenthon et al. proposed to assign keyframes of a variant number according to the activity level of the corresponding scene shot, *see* D. DeMenthon, V. Kobla, and D. Doermann, "Video summarization by curve simplification," Tech. Rep. LAMP-TR-018, Language and Media Processing laboratory, University of Maryland, 1998. The described method represents a video sequence as a trajectory curve in a high dimensional feature space, and uses a recursive binary curve splitting algorithm to find a set of perceptually significant points, which can be used in approximating the video curve. The curve approximation is repeated until the approximation error comes below the user-specified value. Frames corresponding to these perceptually significant points are then used as keyframes to summarize the video contents. Because the curve splitting algorithm assigns more points to segments with larger curvature, this method naturally assigns more keyframes to shots with more variations.

Keyframes extracted from a video sequence may contain duplications and redundancies. For example, in a TV program with two people talking, the video

camera usually switches back and forth between the two persons, and inserts some global views of a scene. Applying the above keyframe selection methods to this video sequence will generate many keyframes that are almost identical. To remove redundancies from the produced keyframes, Yeung et al. selected one keyframe from  
5 each video shot, performed hierarchical clustering on these keyframes based on their visual similarity and temporal distance, and then retained only one keyframe for each cluster, *see* M. Yeung, B. Yeo, W. Wolf, and B. Liu, "Video browsing using clustering and scene transitions on compressed sequences," in *Proc. SPIE on Multimedia Computing and Networking*, vol. 2417, 1995. Girgensohn and Boreczky  
10 also applied the hierarchical clustering technique to group the keyframes into as many clusters as specified by the user. For each cluster, a single keyframe is selected such that the constraints dictated by the requirement of an even distribution of keyframes over the length of the video and a minimum distance between keyframes are met, *see* A. Girgensohn and J. Boreczky, "Time-constrained keyframe selection technique," in  
15 *Proc. IEEE Multimedia Computing and Systems (ICMCS'99)*, 1999.

To create a concise summary of video contents, it is very important to ensure that the summarized representation of the original video (1) contains little redundancy, and (2) gives equal attention to the same amount of contents. While  
20 some of the sophisticated keyframe selection methods address these two issues to variant extents, they often rely on the users to provide either the number of keyframes to be generated, or some thresholds (e.g., a similarity distance between keyframes or approximation errors), which are used in keyframe generation. Accordingly, an optimal set of keyframes can be produced only after several rounds of trials. On the  
other hand, excessive trials could become prohibitive when video images are accessed  
25 through a connection with limited bandwidth, or when a keyframe-set must be created for each video image in a large-scale video database.

Apart from the above problems of keyframe selection, summarizing video contents using keyframes has its own limitations. A video image is a continuous recording of a real-world scene. A set of static keyframes by no means captures the  
30 dynamics and the continuity of the video image. For example, in viewing a movie or

a TV program, the user may well prefer a summarized motion video with a specified time length to a set of static keyframes.

A second important aspect of managing video data is providing tools for efficient searching and retrieval of video sequences according to user-specified queries. It can be a painful task to find either an appropriate video sequence, or the desired portions of the video hidden within a large video data collection. Traditional text indexing and retrieval techniques have turned out to be powerless in indexing and searching video images. To tap into the rich and valuable video resources, video images must be transformed into a medium that is structured, manageable and searchable.

The initial steps toward the aforementioned goal include the segmentation of video sequences into shots for indexing and access, and the extraction of features/metadata from the shots to enable their classification, search, and retrieval. For video shot segmentation, a number of methods have been proposed in past years. Typical video shot segmentation methods include shot segmentation using pixel values, described in K. Otsuji, Y. Tonomura, and Y. Ohba, "Video browsing using brightness data," in *SPIE Proc. Visual Communications and Image Processing*, (Boston), pp. 980-989, 1991, and A. Hampapur, R. Jain, and T. Weymouth, "Digital video segmentation," in *Proceedings of ACM Multimedia 94*, (San Francisco), Oct. 1994. Another video segmentation method, described in H. Ueda, T. Miyatake, and S. Yoshizawa, "Impact: An interactive natural-motion-picture dedicated multimedia authoring system," in *Proc. ACM SIGCHI'91*, (New Orleans), Apr. 1991, relies on global or local histograms. The use of motion vectors in video segmentation is described in H. Ueda, et al., see above. Discrete cosine transform (DCT) coefficients from MPEG files can also be used for video shot segmentation, see F. Arman, A. Hsu, and M.Y. Chiu, "Image processing on encoded video sequences," *Multimedia Systems*, vol. 1, no. 5, pp. 211-219, 1994.

Apart from the aforementioned methods, many other video segmentation techniques have been developed recently. While the vast majority of video segmentation methods use a simple approach of frame-pair comparisons and can

detect only abrupt shot boundaries, some more sophisticated segmentation techniques use additional frames in the aforementioned comparison operation to provide for the detection of gradual scene changes, see H. Zhang, A. Kankanhalli, and S. Smoliar, "Automatic partitioning of full-motion video," Multimedia Systems, vol. 1, pp. 10-28, 5 1993. As it pertains to the video shot retrieval and classification, the most common approach to date has been to first carry out the video shot segmentation, perform additional operations to extract features from each detected shot, and then create indices and metrics using the extracted features to accomplish shot retrieval and classification. In systems based on this described approach, several of the 10 aforementioned processing steps must be performed simultaneously. As a result, these systems usually suffer from high computational costs and long processing times.

Accordingly, there is a recognized need for, and it would be advantageous to have an improved technique that aims to automatically create an optimal and non- 15 redundant summarization of an input video sequence, and to support different user requirements for video browsing and content overview by outputting either the optimal set of keyframes, or a summarized version of the original video with the user-specified time length.

There is also a demand for, and it would be advantageous to have an improved 20 technique for segmenting video sequences into shots for indexing and access, and the extracting features/metadata from the segmented shots to enable their classification, search, and retrieval.

### **SUMMARY OF THE INVENTION**

25 Accordingly, it is one object of the present invention to provide an improved technique for automatically creating an optimal and nonredundant video sequence summarization.

It is another object of the invention to provide a method and system for effective video segmentation and classification.

To achieve the above and other features and realize the benefits and advantages of the invention, there is provided a method and system for video summarization using singular value decomposition. For an input video sequence, the inventive method creates a feature-frame matrix  $A$ , and performs a singular value decomposition thereon. From this singular value decomposition, the present invention derives a refined feature space having fewer dimensions, and a metric to measure the degree of visual changes of each video segment in the input video sequence. In the refined feature space, the content value of a video segment is measured using its degree of visual changes.

For the input video sequence, the inventive method finds the most static video segment, defines it as an information unit, and uses the content value computed from this segment as a threshold to cluster the remaining frames in the video sequence. Using this clustering result, either an optimal set of keyframes, or a summarized motion video with a user-specified time length is generated.

Another aspect of the invention is a method and system for video segmentation and classification. According to the inventive method, a similarity is computed between each of the frames in the input video sequence and a precedent or subsequent frame. The input video sequence is then segmented into a plurality of shots according to the computed similarity values. The similarity metric is defined using the properties of the singular value decomposition. This similarity metric is also used in the application of retrieving visually similar video frames.

In addition to the above similarity metric, a metric to measure the evenness of the color distribution of a frame is also derived from properties of the singular value decomposition to facilitate video classification.

Other aspects of the inventive method include arranging the selected frames into a feature frame matrix, and performing the singular value decomposition on this feature frame matrix. Performing a singular value decomposition also produces a matrix, each column thereof representing a frame in the refined feature space corresponding to a frame in the input video sequence.

In another aspect, features are extracted from each of the shots.

According to the inventive method, the similarity between frames in the input video sequence can be determined using a refined feature space representation of the input video sequence.

Further improvements include comparing the computed similarity to at least  
5 two threshold similarities and segmenting the input video sequence according to the result of this comparison.

### **BRIEF DESCRIPTION OF THE DRAWINGS**

The above and other objects and advantages of the present invention will no  
10 doubt become clear and apparent from the following detailed description of preferred embodiments thereof with reference to the attached drawing, wherein:

Fig. 1 is the block diagram of the inventive video summarization method.

Fig. 2 is the block diagram of the inventive clustering method.

Fig. 3 is the block diagram of the inventive summary composition method.

15 Fig. 4 is the block diagram of the inventive video segmentation and classification method.

Fig. 5 is the block diagram of the inventive shot segmentation method.

### **DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS**

20 The inventive video summarization technique aims to automatically create an optimal and nonredundant video summarization. The invention also seeks to fulfill different user requirements for video browsing and content overview by outputting either an optimal set of keyframes representing the input video sequence, or a summarized motion video of the original video with the user specified time length.

25 The inventive video summarization method uses the aforementioned singular value decomposition as its basic working instrument. Singular value decomposition is known for its capabilities of deriving a low dimensional refined feature space from a high dimensional raw feature space, and of capturing the essential structure of a data set in the feature space. See S. Deerwester, S. Dumais, G. Purnas, T. Landauer,  
30 and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American*

*Society for Information Science*, vol. 41, pp. 391-407, 1990. To reduce the number of frames to be processed by the singular value decomposition, the present invention selects a set of frames that are evenly spaced in the input video sequence (preferably one from every ten frames). For each frame  $i$  in this sampling set, the inventive  
5 technique creates an  $m$ -dimensional feature vector  $A_i$ . Using  $A_i$  as a column, the invention obtains a feature-frame matrix  $\mathbf{A} = [A_1, A_2 \dots A_n]$ . Performing subsequent singular value decomposition on this matrix  $\mathbf{A}$  projects each frame  $i$  from the  $m$ -dimensional raw feature space into a  $\kappa$ -dimensional refined feature space (usually  $\kappa \ll m$ , though this is not required). In this new feature space, noise and trivial  
10 variations in video frames are ignored, and frames with similar color distribution patterns are mapped near to each other. Therefore, the  $\kappa$ -dimensional vectors representing each of the frames in the refined feature space can be used not only for clustering visually similar frames for content summarization, but also for accurately segmenting video frames into shots, and also for similarity matching among the  
15 detected shots.

It will be also appreciated by those of skill in the art that, in the refined feature space, there is a strong correlation between the degree of visual changes in a frame cluster and the locations at which its constituent frames are projected. For many video images, the degree of visual changes is a good indicator of the level of activity  
20 in the images. Taking the video footage only, a static video with almost no changes conveys less information than a dynamic video with abundant changes. Based on the foregoing property of the refined feature space, the content value in a video segment is determined using the locations of its constituent frames in the refined feature space.

Accordingly, in order to summarize the input video according to its content  
25 value, the inventive method first finds the frame cluster in the refined feature space that corresponds to the most static video segment, defines it as an information unit, and uses the content value computed from this frame cluster as a threshold to cluster the rest of the frames in the refined feature space. After the clustering is complete, the inventive method selects a keyframe from each cluster, the selected keyframe  
30 being a frame closest to the center of the cluster. Thus, the inventive approach



ensures that the obtained keyframe set contains little redundancy and gives equal attention to the same amount of contents. To support different user requirements for video browsing and content overview, the inventive system is able to output either the optimal set of keyframes, or a summarized motion video of the original video with the user specified time length.

In addition to the strong correlation between the degree of visual changes in a video segment and the locations in the refined feature space at which the constituent frames of the video segment are projected, a similar correlation exists between the evenness of color distribution in a video frame, and the location of the frame projections. This important property makes singular value decomposition extremely useful for video shot classification. While the degree of visual changes represents the dynamic level of a video segment, the evenness of color distribution reflects its color appearance. The aforementioned properties of singular value decomposition enables the realization of optimal video summarization, accurate video shot segmentation, and effective visual content-based shot classification.

Preferred embodiments of the inventive video summarization and shot segmentation methods will now be described in detail.

### **Exemplary Construction of a Feature Vector**

The video frames of the input video sequence are represented in the method according to an embodiment of the present invention using color histograms. The use of such histograms enables very effective detection of overall differences in image frames. In addition, computations involving histograms have been known to be very cost-effective. This cost-effective property ensures the feasibility and scalability of the inventive method in handling long video sequences.

In one embodiment of the inventive method, three-dimensional histograms in the red-green-blue (RGB) color space are created with five bins for each of R, G, and B, primary colors respectively, resulting in a total of 125 bins. To incorporate the information on the spatial distribution of colors in the video frames, each frame is divided into nine blocks, preferably in a three-by-three manner. The aforementioned three-dimensional color histograms are created for each of the nine blocks. These

nine histograms are then concatenated together to form a 1125-dimensional feature vector for the frame. Using the feature vector of frame  $i$  as the  $i$ 'th column, the inventive method creates the feature-frame matrix  $\mathbf{A}$  representing the input video sequence. Because a small image block does not normally contain a complete color palette, matrix  $\mathbf{A}$  is usually sparse. Therefore, the present invention uses singular value decomposition algorithms for sparse matrices, which are much faster and memory efficient compared to the conventional singular value decomposition algorithms.

Any image features that can be encoded into a vector of a fixed length can be utilized in this inventive method. Alternatively, the feature vector of an image frame can be constructed using a Fourier transformation. As will be undoubtedly appreciated by those of skill in the art, the Fourier feature vector is constructed through a convolution of the image color information using a set of Fourier kernels, for instance sine and cosine. In such case, the coordinates of the feature vector would represent the frequencies of the color distributions within the image.

Yet alternatively, such a feature vector can be constructed using the Wavelet procedure, which is similar to the aforementioned Fourier transformation, except it utilizes a different kernel function, the construction of which is well known in the art.

### Singular Value Decomposition

The inventive video summarization, shot segmentation and classification methods are based on the singular value decomposition, which is defined as follows. Given an  $m \times n$  matrix  $\mathbf{A}$ , where  $m \geq n$ , the singular value decomposition of  $\mathbf{A}$  is defined as:

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \quad (1)$$

where  $\mathbf{U} = [\mathbf{u}_{ij}]$  is an  $m \times n$  column-orthonormal matrix whose columns are called left singular vectors;  $\mathbf{\Sigma} = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_n)$  is an  $n \times n$  diagonal matrix whose diagonal elements are non-negative singular values sorted in descending order, and  $\mathbf{V} = [\mathbf{v}_{ij}]$  is an  $n \times n$  orthonormal matrix whose columns are called right singular vectors.

Additional information on the singular value decomposition can be found in W. Press

et al., *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge, England, Cambridge University Press, 2ed., 1992. If  $\text{rank}(\mathbf{A})=r$ , then  $\Sigma$  satisfies

$$\begin{cases} \sigma_i > 0 : 1 \leq i \leq r \\ \sigma_i = 0 : i > r \end{cases} \quad (2)$$

5        In the inventive video summarization method, applying singular value decomposition to the feature-frame matrix  $\mathbf{A}$  can be interpreted as follows. The singular value decomposition derives a mapping between the  $m$ -dimensional raw feature space occupied by the color histograms and the  $r$ -dimensional refined feature space with all of its axes linearly-independent. Accordingly, the singular value  
10        decomposition transforms each column vector  $i$  of the matrix  $\mathbf{A}$ , which represents the concatenated color histogram of frame  $i$ , into a row vector  $[v_{i1} \ v_{i2} \ \dots \ v_{in}]$  of the matrix  $\mathbf{V}$ . The singular value decomposition also maps each row vector  $j$  of the matrix  $\mathbf{A}$ , which carries the information on the occurrence count of the concatenated histogram entry  $j$  in each of the video frames, into row vector  $[u_{j1} \ u_{j2} \ \dots \ u_{jn}]$  of the matrix  $\mathbf{U}$ .

15        The singular value decomposition requires the number of rows  $m$  of the matrix  $\mathbf{A}$  to be greater than or equal to its number of columns  $n$ . If the number of the selected frames in the input video sequence is greater than the number of elements in each of the concatenated color histograms, the singular value decomposition must be carried out on  $\mathbf{A}^T$ , and consequently, the roles of the matrices  $\mathbf{U}$  and  $\mathbf{V}$ , which have  
20        been set forth above, will be exchanged. For simplicity, and without loss of generality, only the processing of the matrix  $\mathbf{A}$  will be described hereinbelow.

      The singular value decomposition has the following additional important property, which has been widely utilized for text indexing and retrieval. The formal proof of that property can be found in G. Golub and C. Loan, *Matrix Computations*,  
25        Baltimore, Johns-Hopkins, 2 ed., 1989.

**Property 1.** Let the singular value decomposition of matrix  $\mathbf{A}$  be given by Equation (1),  $\mathbf{U} = [\mathbf{U}_1 \mathbf{U}_2 \dots \mathbf{U}_n]$ ,  $\mathbf{V} = [\mathbf{V}_1 \mathbf{V}_2 \dots \mathbf{V}_n]$ , and  $\sigma_1 \geq \sigma_2 \dots \geq \sigma_r > \sigma_{r+1} = \dots = \sigma_n = 0$ . Then, matrix  $\mathbf{A}_\kappa$ , defined below is the closest rank- $\kappa$  matrix to  $\mathbf{A}$  for the Euclidean and Frobenius norms.

$$A_{\kappa} = \sum_{i=1}^{\kappa} U_i \cdot \sigma_i \cdot V_i^T \quad (3)$$

The use of  $\kappa$ -largest singular values to approximate the original matrix with the expression of the Equation (3) has significant implications. Discarding small singular values is equivalent to discarding linearly semi-dependent axes of the feature space. The image features corresponding to the discarded axes are practically nonessential for representing the contents of the images in the video sequence.

On the other hand, the truncated refined feature space captures the most of the important underlying structure of the color histograms and the associated video frames, yet at the same time removes the noise or trivial variations in the video frames. Minor differences between the color histograms will be ignored, and video frames with similar color distribution patterns will be mapped near each other in the  $\kappa$ -dimensional refined feature space. The value of  $\kappa$  is a design parameter. Preferably,  $\kappa = 150$ . Experiments have shown that this value of  $\kappa$  gives satisfactory video summarization results.

#### Video Summarization Based on Singular Value Decomposition

Besides the aforementioned properties, singular value decomposition has the following important feature, which provides a basis for the inventive video summarization system.

**Property 2.** The singular value decomposition of  $\mathbf{A}$  is given by Equation (1), wherein  $\mathbf{A} = [A_1 \dots A_i \dots A_n]$ ,  $\mathbf{V}^T = [\psi_1 \dots \psi_i \dots \psi_n]$ , and  $\psi_i = [v_{i1} \ v_{i2} \ \dots \ v_{in}]^T$ . The distance of  $\psi_i$  to the origin of the refined feature space can be defined as:

$$\|\psi_i\| = \sqrt{\sum_{j=1}^{\text{rank}(\mathbf{A})} v_{ij}^2} \quad (4)$$

If  $\text{rank}(\mathbf{A})=n$ , then, from the orthonormal property of matrix  $\mathbf{V}$ ,  $\|\psi_i\|^2 = 1$ , where  $i = 1, 2, \dots, n$ . If

$$\underbrace{\quad\quad\quad}_k$$

$A' = [A_1 \dots A_i^{(1)} \dots A_i^{(k)} \dots A_n]$  is the matrix obtained by duplicating column vector  $A_i$  in  $A$   $k$  times ( $A_i^{(1)} = \dots = A_i^{(k)} = A_i$ ), and  $V'^T = [\psi'_1 \dots \psi'_1 \dots \psi'_k \dots \psi'_n]$  is the corresponding right singular vector matrix obtained from the singular value decomposition, then  $\|\psi'_j\|^2 = 1/k$ , where  $j = 1, 2, \dots, k$ .

5           The above property indicates that if a column vector  $A_i$  of the matrix  $A$  is linearly-independent, the singular value decomposition operation projects it into a vector  $\psi_i$ , whose distance in the refined feature space defined by Equation (4) equals one. When  $A_i$  has some duplicates  $A_i^{(j)}$ , the distance of the projected vector  $\psi'_j$  decreases. The more duplicates  $A_i$  has, the shorter the distance  $\psi'_j$  has.

10           As it pertains to the video domain, the above property of the singular value decomposition provides that the frames in a static video segment (e.g., segments involving anchorpersons or weather maps) will be projected into the locations which are close to the origin of the refined feature space, while frames in video segments containing a lot of changes (e.g., segments containing moving objects, camera pan
 15   and zoom) will be projected into the points farther from the origin. In other words, the location at which the video segment is projected provides information on the degree of visual changes in the segment.

From the viewpoint of content value, a static video with insubstantial visual changes conveys less information than a dynamic video with lots of changes.
 20   Because the degree of visual changes in a video segment  $S_i$  has a strong correlation with the locations at which its constituent frames are projected in the refined feature space, the following quantity can be used as a measure of the content value contained in cluster (video segment)  $S_i$ :

$$INF(S_i) = \sum_{\psi_i \in S_i} \|\psi_i\|^2 \quad (5)$$

25           The inventive system utilizes the content value defined in accordance with above equation (5) in the following manner. The inventive system first finds a cluster closest to the origin in the refined feature space, which corresponds to the most static video segment in the input video sequence. This most static video segment is

subsequently defined as an information unit and the content value computed from the segment is used as a threshold to cluster the rest of the frames in the input video sequence. Thus, the inventive approach ensures that the summarized representation of the original video contains little redundancy and gives equal attention to the same amount of contents.

Accordingly, an embodiment of the inventive video summarization method comprises the following major processing steps, as shown in Fig. 1:

Step 101. Select frames with a fixed interval (preferably a 10-frame interval) from the input video sequence, and create the feature-frame matrix  $\mathbf{A}$  using these selected frames.

Step 102. Perform singular value decomposition on the matrix  $\mathbf{A}$  to obtain the matrix  $\mathbf{V}^T$ , each column vector  $\psi_i$  of this matrix representing frame  $i$  in the refined feature space.

Step 103. In the refined feature space, find the most static cluster, compute the content value of this cluster using Equation (5), and use this value as a threshold to cluster the rest of the frames in the input video sequence.

Step 104. For each obtained cluster  $S_i$  find the longest video shot  $\Theta_i$  contained in the cluster. Discard the cluster whose  $\Theta_i$  is shorter than one second.

Step 105. According to the user's request, output either a set of keyframes, each keyframe representing a video cluster, or a summarized motion video with the user specified time length.

As will be appreciated by those of skill in the art, in Step 103 of the above procedure, finding the most static cluster is equivalent to finding a cluster closest to the origin of the refined feature space. With reference to the used notations, the entire clustering process can be described as follows, with reference to Fig. 2:

Step 201. In the refined feature space, sort all the vectors  $\psi_i$  in ascending order using the distance defined by Equation (4). Initialize all the vectors as unclustered vectors, and set the cluster counter  $C = 1$ .

Step 202. Among the unclustered vectors, find a vector closest to the origin of the refined feature space, and select this vector as a seed vector to form

cluster  $S_c$ . Set the average internal distance of the cluster  $R(S_c) = 0$ , and the frame count  $P_c = 1$ .

Step203. For each unclustered vector  $\psi_i$ , calculate its minimum distance to the cluster  $S_c$ , which is defined as:

5

$$d_{min}(\psi_i, S_c) = \min_{\psi_k \in S_c} D(\psi_i, \psi_k) \quad (6)$$

wherein  $D(\psi_i, \psi_k)$  is defined as the Euclidean distance weighted by the aforementioned singular values. Using the notation associated with Equation (1), the aforementioned Euclidean distance is:

10

$$D(\psi_i, \psi_k) = \sqrt{\sum_{j=1}^{\kappa} \sigma_j (v_{ij} - v_{kj})^2}, \quad (7)$$

wherein  $\kappa$  is the reduced dimension of the refined feature space.

15

In Step 204 if cluster counter  $C = 1$ , go to Step 205 below; otherwise, go to Step 207 below.

In Step 206 add frame  $\psi_i$  to cluster  $S_1$  if, in Step 205

$$R(S_1) = 0 \quad \text{or}$$

$$d_{min}(\psi_i, S_1)/R(S_1) < 5.0$$

20

In Step 208, add frame  $\psi_i$  to cluster  $S_c$  if, in Step 207

$$R(S_c) = 0 \quad \text{or}$$

$$INF(S_c) < INF(S_1) \quad \text{or}$$

$$d_{min}(\psi_i, S_c)/R(S_c) < 2.0$$

25

If frame  $\psi_i$  is added to cluster  $S_c$ , increment frame count  $P_c$  by one, update the content value  $INF(S_c)$  using Equation (5), and update  $R(S_c)$  as follows:

$$R(S_c) = \frac{(P_c - 1)R(S_c) + d_{min}(\psi_i, S_c)}{P_c} \quad (8)$$

Step 209. If there exist unclustered points, increment the cluster counter  $C$  by one, and go to Step 202; otherwise, terminate the operation.

It should be noted that in the above operations, different conditions are used for growing the first and the rest of clusters. The first cluster relies on the distance variation  $d_{min}(\psi_i, S_1)/R(S_1)$  as its growing condition, while the remaining clusters examine the content value as well as the distance variation in their growing process. Condition 2 in Step 207 ensures that the cluster under processing contains the same amount of information as the first cluster, while Condition 3 prevents two frames which are very close to each other from being separated into different clusters. With Condition 2, some long video shots with large visual variations may be clustered into more than one cluster, and consequently, more than one keyframe will be assigned to these types of shots. On the other hand, with the combination of Condition 2 and 3, video shots with very similar visual contents will be clustered together, and only one keyframe will be assigned to this group of video shots. The above features of the inventive clustering method provide substantial advantages in comparison with existing clustering techniques.

In addition, Step 105 of the described summarization process provides for another unique feature of the inventive system. In particular, the inventive system is capable of outputting either an optimal set of keyframes, or a summarized version of the original video having a user-specified time length. When the keyframe output mode is selected by the user, the inventive system performs the singular value decomposition and clustering operations described above. From each obtained cluster, the system selects a frame whose feature vector is the closest to the center of the cluster and designates the selected frame as a keyframe.

The output of a summarized video requires more operations. The inventive system composes a summarized video according to two user-specified parameters: the time length of the summarized video  $T_{len}$ , and the minimum display time of each shot in the summarized video  $T_{min}$ . The process consists of the following main operations, as described with reference to Fig. 3:



Step 301. Let  $C$  be the number of clusters obtained from the above clustering process, and  $N = T_{len}/T_{min}$ . For each cluster  $S_i$ , find the longest video shot  $\Theta_i$ .

Step 302. If  $C \leq N$ , go to Step 303 below; otherwise, go to Step 304 below.

5 Step 303. Select all the shots  $\Theta_i$  wherein  $i = 1, 2, \dots, C$ , and assign an equal time length  $T_{len}/C$  to each of the shots.

Step 304. Sort shots  $\Theta_i$  in the descending order by their length, select the top  $N$  shots, and assign an equal time length  $T_{min}$  to each selected shot.

10 Step 305. Sort the selected shots by the time code, based on this sorted order, get from each selected shot a portion of the assigned time length, and insert that portion into the summarized video.

Given the user-specified parameters  $T_{len}$  and  $T_{min}$ , the maximum number of video shots, which can be included in the summarized video equals  $N = T_{len}/T_{min}$ . If the total number of shots  $C \leq N$ , then all the shots will be assigned a slot in the summarized video (Step 304); otherwise, the shots will be selected in descending order of their lengths to fill the summarized video (Step 304). Here, the parameter  $T_{min}$  can be considered as a control knob for the user to select between depth-centric and breadth-centric summarization. A small value of  $T_{min}$  will produce a breadth-centric video summary, which consists of a larger number of shots with shorter lengths, while a large value for  $T_{min}$  will produce a depth-centric video summary consisting of fewer shots, each shot being longer in length. Moreover, because the clustering process is performed such that all the resultant clusters contain approximately the same amount of information, it is natural to assign the same time length to each selected shot in the summarized video.

25 The inventive video summarization system was implemented using C++ programming language and evaluated using a wide variety of input video sequences. The input video sequences used in the testing of the inventive system included news reports, documentary, political debates, and live coverage of various events. Each test video sequence lasted between 5 and 30 minutes. In one example, a summary of  
30 a 5-minute video documentary created by the inventive video summarization system,



$$\text{SIM}(i, j) = D(\psi_i, \psi_j) = \sqrt{\sum_{l=1}^{\kappa} \sigma_l (v_{il} - v_{jl})^2}, \quad (9)$$

wherein  $\psi_i, \psi_j$  are vectors representing frames  $i$  and  $j$  in the refined feature space, respectively, and  $\sigma_l$ 's are the singular values obtained in the singular value decomposition.

- 5 In addition to the aforementioned important features, the singular value decomposition has the following additional property:

**Property 3.** The singular value decomposition of  $\mathbf{A}$  is given by Equation (1), wherein  $\mathbf{A} = [A_1 \dots A_i \dots A_n]$ ,  $\mathbf{V}^T = [\psi_1 \dots \psi_i \dots \psi_n]$ ,  $A_i = [a_{1i} \ a_{2i} \ \dots \ a_{mi}]^T$ , and  $\psi_i = [v_{i1} \ v_{i2} \ \dots \ v_{in}]^T$ . Define the singular value weighted length of  $\psi_i$  as:

$$10 \quad \|\psi_i\|_{\Sigma} = \sqrt{\sum_{j=1}^{\text{rank}(\mathbf{A})} \sigma_j^2 v_{ij}^2}.$$

where  $\sigma_j$ 's are the singular values. (10)

$$\text{Then, } \|\psi_i\|_{\Sigma}^2 = A_i \cdot A_i = \sum_{j=1}^m a_{ji}^2.$$

- Property 3 can be used as an indicator of the evenness of color distribution in frames and shots. Because  $A_i$  is the concatenated histogram vector of frame  $i$ , the sum of its elements  $a_{ji}$  is a constant  $C$  (which is equal to the number of pixels in the frame). Therefore,  $\|\psi_i\|_{\Sigma}^2$  reaches its minimum when  $a_{1i} = a_{2i} = \dots = a_{mi}$ , and it reaches its maximum when one of its elements  $a_{ki}$  is equal to  $C$  and the remaining elements are all equal to zero. Accordingly, the singular value weighted length  $\|\psi_i\|_{\Sigma}^2$  is proportional to the evenness of the color distribution of the corresponding frame  $i$ . This length becomes the shortest when substantially all colors are present in the frame  $i$  in substantially equal amounts (i.e. when frame  $i$  has substantially even color distribution), and it becomes the longest when the frame  $i$  contains only one color.

- A preferred embodiment of the inventive video segmentation and classification method comprises the following major steps, as depicted in Fig. 4:

Step 401. Sample the input video sequence with a fixed rate, preferably 10 frames per second, and create the feature-frame matrix  $\mathbf{A}$  as described above.

Step 402. Perform singular value decomposition on the matrix  $\mathbf{A}$  to obtain matrices  $\Sigma = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_n)$  and  $\mathbf{V}^T = [\psi_1 \dots \psi_i \dots \psi_n]$ .

5 Step 403. Compute the similarity  $\text{SIM}(i, i + 1)$  defined by Equation (9) for all the frames in the sample set, and segment the video sequence into video shots along the time axis (see the following segmentation algorithm for the detail).

Step 404. For each video shot  $\Theta_s$ , compute the following two average lengths:

$$10 \quad \overline{\|\Theta_s\|}^2 = \frac{1}{P(\Theta_s)} \cdot \sum_i \|\psi_i\|^2 \quad (11)$$

$$\overline{\|\Theta_s\|}^2_\Sigma = \frac{1}{P(\Theta_s)} \cdot \sum_i \|\psi_i\|_\Sigma^2, \quad (12)$$

wherein  $\psi_i \in \Theta_s$ , and  $P(\Theta_s)$  is the number of frames included in shot  $\Theta_s$ .

The above two values indicate the degree of visual changes, and the evenness of color distributions in the shot  $\Theta_s$ , respectively.

15 Step 405. Compute the average feature vector  $\Psi_s$  for each shot  $\Theta_s$ . Distance  $D(\Psi_X, \Psi_Y)$  defines the visual similarity between shots  $\Theta_X$  and  $\Theta_Y$ .

In the above embodiment of the inventive algorithm, Steps 401 and 402 perform singular value decomposition for the shot segmentation; Step 403 conducts the shot segmentation itself; and Steps 404 and 405 compute the metrics for each  
20 detected shot to enable the assessment of the color distributions, dynamic levels, and visual similarities of all the shots in the video sequence.

The step of shot segmentation (Step 403) involves two thresholds,  $T_{low}$  and  $T_{high}$ . If the distance between two consecutive frames is below  $T_{low}$ , the two frames will be grouped into the same shot without further examination. If this distance is  
25 above  $T_{high}$ , a shot boundary will be declared. If this distance is between  $T_{low}$  and  $T_{high}$ , further examination involving more video frames will be performed to determine if the detected large distance is caused by the temporary variation, or a gradual scene transition. The following algorithm describes the implementation of

the preferred embodiment of the inventive shot segmentation method, as depicted in Fig. 5:

Step 501. Set shot counter  $S = 1$ , and frame index  $I = 1$ .

Step 502. Create shot  $\Theta_s$  with frame  $I$  as its first element.

5 Step 503. If  $D(\psi_I, \psi_{I+1}) \leq T_{low}$ , then insert frame  $I + 1$  into shot  $\Theta_s$  and increment  $I$  by one (Step 504). Repeat this step if  $I$  is not the last frame (Step 505); otherwise, go to Step 514.

Step 506. If  $D(\psi_I, \psi_{I+1}) > T_{high}$ , mark the location between frames  $I$  and  $I + 1$  as a shot boundary and increment  $S$  and  $I$  by one. Then, return to Step 502.

10 If  $T_{low} < D(\psi_I, \psi_{I+1}) \leq T_{high}$ , do the following:

Step 507. Find a frame  $X > I + 1$  which satisfies the condition  $D(\psi_I, \psi_{I+1}) \leq T_{low}$  (Step 507).

Step 510. If  $D(\psi_X, \psi_I) > T_{high}$ , mark the frames between  $I + 1$  and  $X$  as a gradual transition between the two scene shots (Step 508); set  $I = X + 1$ , and increment the shot counter  $S$  by one (Step 508). Then, go to Step 514.

15 If  $D(\psi_I, \psi_{I+1}) \leq T_{high}$ , group frames from  $I + 1$  to  $X$  into shot  $\Theta_s$ , and set  $I = X + 1$  (Step 513). Then, return to Step 503.

Step 514. If the last frame has been reached, terminate the entire operation; otherwise, return to Step 502.

20 Table 1: Evaluation of a Preferred Embodiment of the Inventive Video Segmentation and Classification System

	Abrupt Shot Cut		Gradual Transition		Shot Classification	
	Recall	Precision	Recall	Precision	Recall	Precision
Local Histogram Method	92.6%	72.1%	-	-	-	-
Inventive Method	97.3%	92.7%	94.3%	87.0%	90.2%	85.1%

A preferred embodiment of the inventive video shot segmentation and classification system was evaluated using a total of two hours of CNN news video programs. The used video footage contained almost all possible video editing effects

such as abrupt scene changes, fades, wipes, dissolves, etc. The used footage also contained a great variety of scene categories such as portraits, landscapes, interviews, crowds, moving camera/objects, etc. For the sake of comparison, a local histogram-based shot segmentation method was also implemented and evaluated using the same  
5 set of video programs. The aforementioned local histogram method was chosen for comparison with the inventive method because its performance was reported to be one of the best among the existing segmentation techniques. See J. Boreczky and L. Rowe, "Comparison of video shot boundary detection techniques," in *Proceedings of SPIE: Storage and Retrieval for Image and Video Databases IV*, vol. 2670, 1996.

10 The experimental results are listed in Table 1.

Persons of skill in the art will undoubtedly appreciate that for abrupt shot cut detection, the inventive system provides a remarkable improvement in recall and precision over the conventional technique. Such dramatic improvements are achieved because of the frame comparison in the truncated feature space derived from the  
15 singular value decomposition, and the use of the two thresholds  $T_{high}$  and  $T_{low}$  which divide the entire domain of the frame distance into the low, gray, and high zones. As set forth above, if the distance between two consecutive frames falls into the gray zone, more frames will be examined to determine if the large distance is due to the presence of video noise, jitters from camera/object motions, or the genuine scene  
20 change. This inventive approach greatly reduces outliers and results in a high recall, high precision rates of the shot boundary detection. As well known to persons of ordinary skill in the art, the term outliers refers to the detected scene changes caused not by the changes in the video content, but by secondary effects such as camera jitter.

25 In addition, the inventive system is also capable of detecting gradual scene transitions, and classifying the detected shots into four categories such as identical shots, shots with high degree variations, static shots without remarkable changes, and shots with a uniform color (e.g., black/white frames). In many video programs, the same persons or the same scenes appear repeatedly (e.g. anchorpersons,  
30 interviewers/interviewees). Finding these identical scenes is essential for detecting

and eliminating duplicates and redundancies, which is critical for generating concise video content summaries. On the other hand, dynamic shots with abundant variations may contain either camera pans and zooms, which aim at capturing the entire event, or dramatic object motions, which come from highly dynamic scenes. The ability to  
5 identify dynamic shots is extremely important to achieving the ultimate goal of detecting visually important scenes.

Finally, because black or white frames often appear around scene shot boundaries, for example right before or right after TV commercials, detecting these kinds of frames is useful for many applications. In order to conserve the memory  
10 resources, the recall and precision values for shot classification are obtained by averaging the recall and precision values of the aforementioned four shot categories.

From the above table, it will become clear to those of skill in the art that the inventive system has achieved a competitive performance in the gradual scene transition detection as well as the shot classification. In contrast to many traditional  
15 shot classification systems, which rely heavily on heuristic rules and sets of thresholds, the inventive system classifies shots based on the metrics derived from the properties of the singular value decomposition. This feature of the present invention provides for simple, robust, and accurate classification of video shots.

Accordingly, it will be readily appreciated by persons of skill in the art that  
20 the inventive shot segmentation and classification system based on singular value decomposition successfully achieves the aforementioned goals of accurate video shot segmentation and visual content-based shot classification.

While the invention has been described herein using preferred embodiments thereof, it will be readily appreciated by those skilled in the art that various  
25 modifications in form and detail may be made therein without departing from the scope and spirit of the invention, as defined in and by the appended claims.

WHAT IS CLAIMED IS:

- 1           1.     A method for summarizing a content of an input video sequence, said  
2 input video sequence comprising a plurality of frames, said plurality of frames being  
3 grouped into a plurality of video segments, said method comprising:

- 4 (a) selecting a frame cluster in said input video sequence which  
5 corresponds to a most static one of said video segments;  
6 (b) computing a content value in said selected frame cluster;  
7 (c) using said computed content value to cluster remaining frames  
8 in said input video sequence.

1 2. The method of claim 1, wherein in said (a) said frame cluster is  
2 selected using a refined feature space representation of said input video sequence.

1 3. The method of claim 1, wherein in said (a) each of said plurality of  
2 frames is transformed into a histogram vector indicative of a spatial distribution of  
3 colors in said each of said plurality of frames.

1 4. The method of claim 3, wherein in said (a) each of said plurality of  
2 frames is divided into a plurality of blocks, each of said plurality of blocks being  
3 represented by a histogram in a color space indicative of a distribution of colors  
4 within each of said plurality of blocks.

1 5. The method of claim 3, wherein each of said plurality of frames is  
2 divided into a plurality of blocks and each said histogram vector comprises a plurality  
3 of histograms in a color space, each of said plurality of histograms corresponding to  
4 one of said plurality of blocks.

1 6. The method of claim 2, wherein said refined feature space  
2 representation is obtained using a singular value decomposition of said input video  
3 sequence.

1 7. The method of claim 6, wherein said singular value decomposition is  
2 performed using frames selected with a fixed interval from said input video sequence.

1 8. The method of claim 7, wherein said selected frames are arranged into  
2 a feature frame matrix, and wherein said singular value decomposition is performed  
3 on said feature frame matrix.



1           9.     The method of claim 6, wherein said singular value decomposition  
2 produces a matrix, each column of said matrix representing a frame in a refined  
3 feature space corresponding to a frame in said input video sequence.

1           10.    The method of claim 1, further comprising (d) using said clustered  
2 frames to output a motion video representative of a summary of said input video  
3 sequence.

1           11.    The method of claim 1, further comprising (d) outputting a plurality of  
2 keyframes, each of said plurality of keyframes representative of said clustered frames.

1           12.    The method of claim 2, wherein said selecting comprises locating a  
2 cluster closest to an origin of said refined feature space.

1           13.    The method of claim 2, wherein said (c) comprises:

2                   (c)(1) sorting a plurality of vectors in said refined feature space in  
3                   ascending order according to a distance of each of said vectors  
4                   to an origin of said refined feature space representation;

5                   (c)(2) selecting a vector among said sorted vectors which is closest to  
6                   an origin of said refined feature space representation and  
7                   including said selected vector into a first cluster;

8                   (c)(3) clustering said plurality of sorted vectors in said refined feature  
9                   into a plurality of clusters according to a distance between each  
10                  of said plurality of sorted vectors and vectors in each of said  
11                  plurality of clusters and an amount of information in each of  
12                  said plurality of clusters.

1           14.    The method of claim 13, wherein in said (c)(3) said plurality of sorted  
2 vectors are clustered into said plurality of clusters such that said amount of  
3 information in each of said plurality of clusters does not exceed an amount of  
4 information in said first cluster.

1           15.    The method of claim 13, wherein said first cluster is composed of  
2 frames based on a distance variation between said frames and an average distance  
3 between frames in said first cluster.

1           16.    The method of claim 13, wherein each of said plurality of clusters is  
2 composed of frames based on a distance variation between said frames and an  
3 average distance between frames in said each of said plurality of clusters.

1           17.    A method for summarizing a content of an input video sequence, said  
2 method comprising:

- 3                   (a)    selecting frames from said input video sequence, said selected  
4 frames being taken at a fixed interval;
- 5                   (b)    creating a feature frame matrix using said selected frames;
- 6                   (c)    performing a singular value decomposition on said feature  
7 frame matrix to obtain a matrix representing said video  
8 sequence in a refined feature space;
- 9                   (d)    selecting a cluster in said refined feature space corresponding  
10 to a most static video segment;
- 11                  (e)    computing a content value corresponding to said selected  
12 cluster;
- 13                  (f)    using said computed content value to cluster frames in said  
14 input video sequence.

1           18.    A method for segmenting an input video sequence, said input video  
2 sequence comprising a plurality of frames, said plurality of frames being grouped  
3 into a plurality of video shots, said method comprising:

- 4                   (a)    computing a similarity between each of said plurality of frames  
5 and a frame preceding said each of said plurality of frames in  
6 time;
- 7                   (b)    segmenting said input video sequence into said plurality of  
8 video shots according to said computed similarity.

1           19.     The method of claim 18, wherein said similarity is calculated using a  
2 refined feature space representation of said input video sequence.

1           20.     The method of claim 19, wherein said refined feature space  
2 representation is created using a singular value decomposition of said input video  
3 sequence.

1           21.     The method of claim 20, wherein said singular value decomposition is  
2 performed using frames selected with a fixed interval from said input video sequence.

1           22.     The method of claim 21, wherein said selected frames are arranged  
2 into a feature frame matrix, and wherein said singular value decomposition is  
3 performed on said feature frame matrix.

1           23.     The method of claim 22, wherein said performed singular value  
2 decomposition produces a matrix, each column of said produced matrix comprising a  
3 frame in said refined feature space representing a frame in said input video sequence.

1           24.     The method of claim 18, further comprising (c) extracting features  
2 from each of said plurality of video shots.

3           25.     A method for determining a similarity between a first and a second  
4 frame in an input video sequence, said method comprising:

- 5                   (a)     calculating a refined feature space representation of said input
- 6                             video sequence;
- 7                   (b)     using said calculated representation to compute said similarity
- 8                             between said first and said second frames.

9           26.     The method of claim 25, wherein in said (a) said refined feature space  
10 representation is calculated using a singular value decomposition.



1           33.    The computer-readable medium of claim 31, wherein in said (a) each  
2 of said plurality of frames is transformed into a histogram vector indicative of a  
3 spatial distribution of colors in said each of said plurality of frames.

1           34.    The computer-readable medium of claim 33, wherein in said (a) each  
2 of said plurality of frames is divided into a plurality of blocks, each of said plurality  
3 of blocks being represented by a histogram in a color space indicative of a  
4 distribution of colors within each of said plurality of blocks.

1           35.    The computer-readable medium of claim 33, wherein each of said  
2 plurality of frames is divided into a plurality of blocks and each said histogram vector  
3 comprises a plurality of histograms in a color space, each of said plurality of  
4 histograms corresponding to one of said plurality of blocks.

1           36.    The computer-readable medium of claim 32, wherein said refined  
2 feature space representation is obtained using a singular value decomposition of said  
3 input video sequence.

1           37.    The computer-readable medium of claim 36, wherein said singular  
2 value decomposition is performed using frames selected with a fixed interval from  
3 said input video sequence.

1           38.    The computer-readable medium of claim 37, wherein said selected  
2 frames are arranged into a feature frame matrix, and wherein said singular value  
3 decomposition is performed on said feature frame matrix.

1           39.    The computer-readable medium of claim 33, wherein said singular  
2 value decomposition produces a matrix, each column of said matrix representing a  
3 frame in a refined feature space corresponding to a frame in said input video  
4 sequence.

1           40.    The computer-readable medium of claim 31, further comprising (d)  
2    using said clustered frames to output a video representative of a summary of said  
3    input video sequence.

1           41.    The computer-readable medium of claim 31, further comprising (d)  
2    outputting a plurality of keyframes, each of said plurality of keyframes representative  
3    of said clustered frames.

1           42.    The computer-readable medium of claim 32, wherein said selecting  
2    comprises locating a cluster closest to an origin of said refined feature space.

1           43.    The computer-readable medium of claim 32, wherein said (c)  
2    comprises:

- 3                   (1)    sorting a plurality of vectors in said refined feature space in  
4                            ascending order according to a distance of each of said vectors  
5                            to an origin of said refined feature space;  
6                   (2)    selecting a vector among said sorted vectors which is closest to  
7                            an origin of said refined feature space and including said  
8                            selected vector into a first cluster;  
9                   (3)    clustering said plurality of sorted vectors in said refined feature  
10                           into a plurality of clusters according to a distance between each  
11                           of said plurality of sorted vectors and each of said plurality of  
12                           clusters and an amount of information in each of said plurality  
13                           of clusters.

1           44.    The computer-readable medium of claim 38, wherein in said (3) said  
2    plurality of sorted vectors are clustered into said plurality of clusters such that said  
3    amount of information in each of said plurality of clusters does not exceed an amount  
4    of information in said first cluster.

1           45.     The computer-readable medium of claim 38, wherein said first cluster  
2 is composed of frames based on a distance variation between said frames and said  
3 first cluster.

1           46.     The computer-readable medium of claim 38, wherein each of said  
2 plurality of clusters is composed of frames based on a distance variation between said  
3 frames and said each of said plurality of clusters.

1           47.     A computer-readable medium containing a program for summarizing a  
2 content of an input video sequence, said program comprising:

- 3                   (a)     selecting frames with a fixed interval from said input video  
4                             sequence;  
5                   (b)     creating a feature frame matrix using said selected frames;  
6                   (c)     performing a singular value decomposition on said feature  
7                             frame matrix to obtain matrix representing said video sequence  
8                             in refined feature space;  
9                   (d)     selecting a cluster in said refined feature space corresponding  
10                            to a most static video segment;  
11                   (e)     computing a content value corresponding to said selected  
12                            cluster;  
13                   (f)     using said computed content value to cluster frames in said  
14                            input video sequence.

1           48.     A computer-readable medium containing a program for segmenting an  
2 input video sequence, said input video sequence comprising a plurality of frames, said  
3 plurality of frames being grouped into a plurality of video shots, said program  
4 comprising:

- 5                   (a)     computing a similarity between each of said plurality of frames  
6                             and a subsequent in time frame;  
7                   (b)     segmenting said input video sequence into a plurality of shots  
8                             according to said computed similarity.

1           49.    The computer-readable medium of claim 18, wherein said similarity is  
2   calculated using a refined feature space representation of said input video sequence.

1           50.    The computer-readable medium of claim 19, wherein said refined  
2   feature space representation is created using a singular value decomposition of said  
3   input video sequence.

1           51.    The computer-readable medium of claim 20, wherein said singular  
2   value decomposition is performed using frames selected with a fixed interval from  
3   said input video sequence.

1           52.    The computer-readable medium of claim 21, wherein said selected  
2   frames are arranged into a feature frame matrix, and wherein said singular value  
3   decomposition is performed on said feature frame matrix.

1           53.    The computer-readable medium of claim 22, wherein said performed  
2   singular value decomposition produces a matrix, each column of said produced  
3   matrix comprising a frame in said refined feature space representing a frame in said  
4   input video sequence.

1           54.    The computer-readable medium of claim 18, wherein said program  
2   further comprises (c) extracting features from each of said plurality of video shots.

3           55.    A computer-readable medium containing a program for determining a  
4   similarity between a first and a second frames in an input video sequence, said  
5   program comprising:

- 6           (a)    calculating a refined feature space representation of said input video  
7                   sequence; and
- 8           (b)    using said calculated representation to compute said similarity between  
9                   said first and said second frames.



10           56.    The computer-readable medium of claim 25, wherein in said (a) said  
11 refined feature space representation is calculated using a singular value  
12 decomposition.

1           57.    The computer-readable medium of claim 18, wherein in said (b) said  
2 computed similarity is compared to at least two threshold similarities, and said input  
3 video sequence is segmented according to a result of said comparison.

1           58.    The computer-readable medium of claim 48, wherein if in said (b) said  
2 computed similarity is below a first threshold similarity, said each of said plurality of  
3 frames is put into a one of said plurality of video shots containing said precedent in  
4 time frame.

5           59.    The computer-readable medium of claim 48, wherein if in said (b) said  
6 computed similarity is above a second threshold similarity, said each of said plurality  
7 of frames is designated as a shot boundary.

8           60.    The computer-readable medium of claim 48, wherein if in said (b) said  
9 computed similarity is between a first threshold similarity and a second threshold  
10 similarity, said each of said plurality of frames is put into a one of said plurality of  
11 video shots according to a further analysis performed using additional frames from  
12 said plurality of frames.

1           61.    The method of claim 18, further comprising (c) extracting features  
2 from each of said plurality of video shots and using said extracted features to index  
3 said plurality of video shots.

1           62.    The method of claim 61, wherein said extracted features are features of  
2 a video frame representative of said each of said plurality of video shots.

1           63.    The computer-readable medium of claim 48, wherein said program  
2 further comprises (c) extracting features from each of said plurality of video shots and  
3 using said extracted features to index said plurality of video shots.

1           64.    The method of claim 63, wherein said extracted features are features of  
2 a video frame representative of said each of said plurality of video shots.

1           65.    A method of calculating a degree of visual changes in a video shot,  
2 said video shot comprising a plurality of frames, said method comprising:

- 3           (a)    performing a singular value decomposition on said plurality of frames,  
4                   wherein said singular value decomposition produces a matrix, each  
5                   column of said matrix representing a frame in a refined feature space  
6                   corresponding to a frame in said plurality of frames;  
7           (b)    using said matrix to calculate said degree of visual changes in said  
8 video shot.

1           66.    The method of claim 65, wherein said (b) comprises calculating said  
2 degree of visual changes in said video shot as a sum  $\sqrt{\sum_{j=1}^{\text{rank}(A)} V_{ij}^2}$ , wherein  $v_{ij}$  are  
3 elements of said matrix.

1           67.    A computer-readable medium containing a program for calculating a  
2 degree of visual changes in a video shot, said video shot comprising a plurality of  
3 frames, said program comprising:

- 4           (a)    performing a singular value decomposition on said plurality of frames,  
5                   wherein said singular value decomposition produces a matrix, each  
6                   column of said matrix representing a frame in a refined feature space  
7                   corresponding to a frame in said plurality of frames;  
8           (b)    using said matrix to calculate said degree of visual changes in said  
9 video shot.

1           68.    The computer-readable medium of claim 67, wherein said (b)  
2 comprises calculating said degree of visual changes in said video shot as a sum

3  $\sqrt{\sum_{j=1}^{\text{rank}(A)} V_{ij}^2}$ , wherein  $v_{ij}$  are elements of said matrix.

69. A method of calculating an evenness of color distributions in a video shot, said video shot comprising a plurality of frames, said method comprising:

- (a) performing a singular value decomposition on said plurality of frames, wherein said singular value decomposition produces a matrix, each column of said matrix representing a frame in a refined feature space corresponding to a frame in said plurality of frames;
- (b) using said matrix to calculate said evenness of color distribution in said video shot.

70. The method of claim 69, wherein said (b) comprises calculating said evenness of color distribution in said video shot as a sum  $\sqrt{\sum_{j=1}^{\text{rank}(A)} \sigma_j^2 v_{ij}^2}$ , wherein said  $v_{ij}$  are elements of said matrix and said  $\sigma_j$  are singular values obtained in said singular value decomposition.

71. A computer-readable medium containing a program for calculating an evenness of color distributions in a video shot, said video shot comprising a plurality of frames, said method comprising:

- (a) performing a singular value decomposition on said plurality of frames, wherein said singular value decomposition produces a matrix, each column of said matrix representing a frame in a refined feature space corresponding to a frame in said plurality of frames;
- (b) using said matrix to calculate said evenness of color distribution in said video shot.

72. The computer readable medium of claim 71, wherein said (b) comprises calculating said evenness of color distribution in said video shot as a sum

$\sqrt{\sum_{j=1}^{\text{rank}(A)} \sigma_j^2 v_{ij}^2}$ , wherein said  $v_{ij}$  are elements of said matrix and said  $\sigma_j$  are singular values obtained in said singular value decomposition.

## **ABSTRACT OF THE DISCLOSURE**

In a technique for video segmentation, classification and summarization based on the singular value decomposition, frames of the input video sequence are represented by vectors composed of concatenated histograms descriptive of the spatial distributions of colors within the video frames. The singular value decomposition maps these vectors into a refined feature space. In the refined feature space produced by the singular value decomposition, the invention uses a metric to measure the amount of information contained in each video shot of the input video sequence. The most static video shot is defined as an information unit, and the content value computed from this shot is used as a threshold to cluster the remaining frames. The clustered frames are displayed using a set of static keyframes or a summary video sequence. The video segmentation technique relies on the distance between the frames in the refined feature space to calculate the similarity between frames in the input video sequence. The input video sequence is segmented based on the values of the calculated similarities. Finally, average video attribute values in each segment are used in classifying the segments.

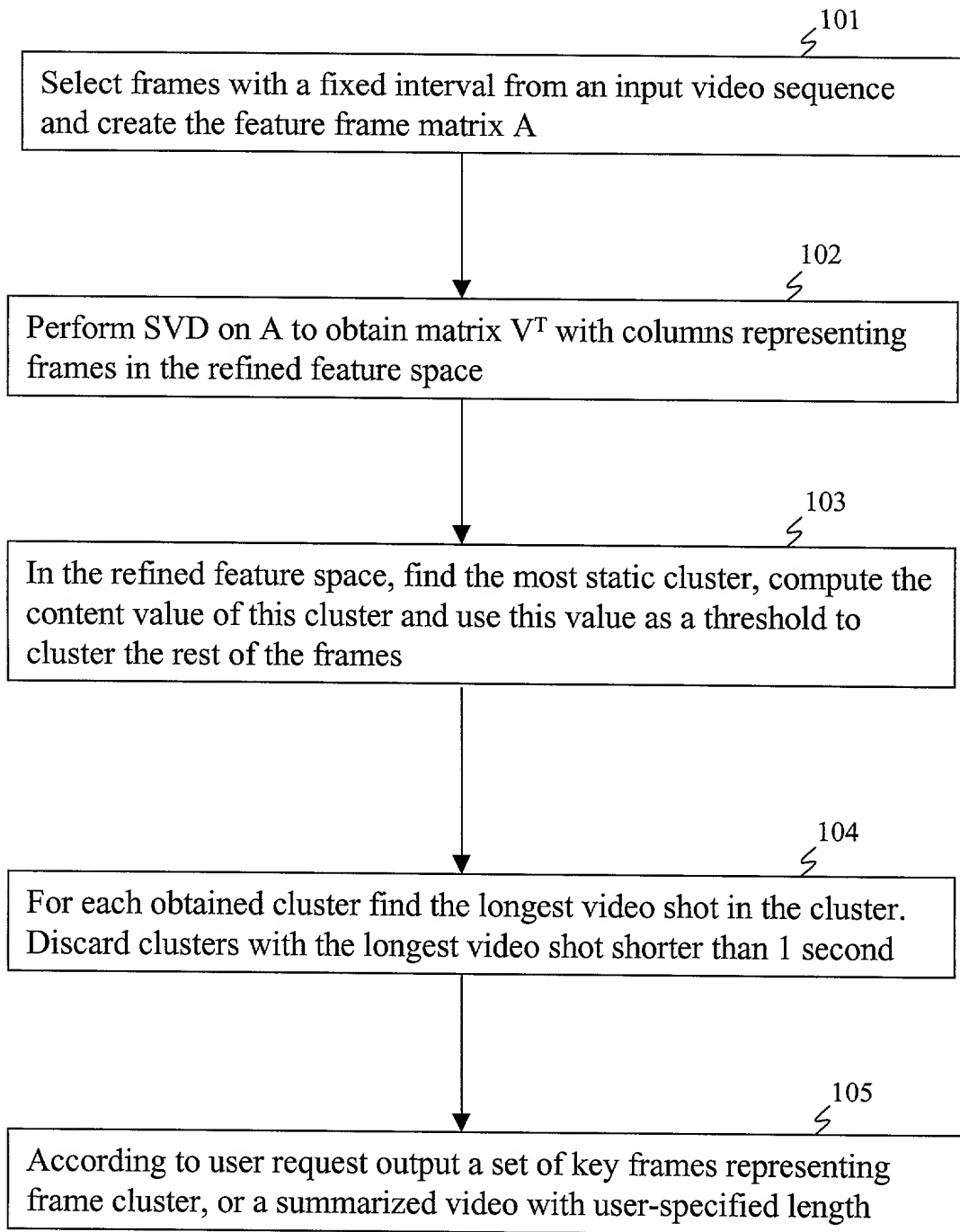


FIG. 1

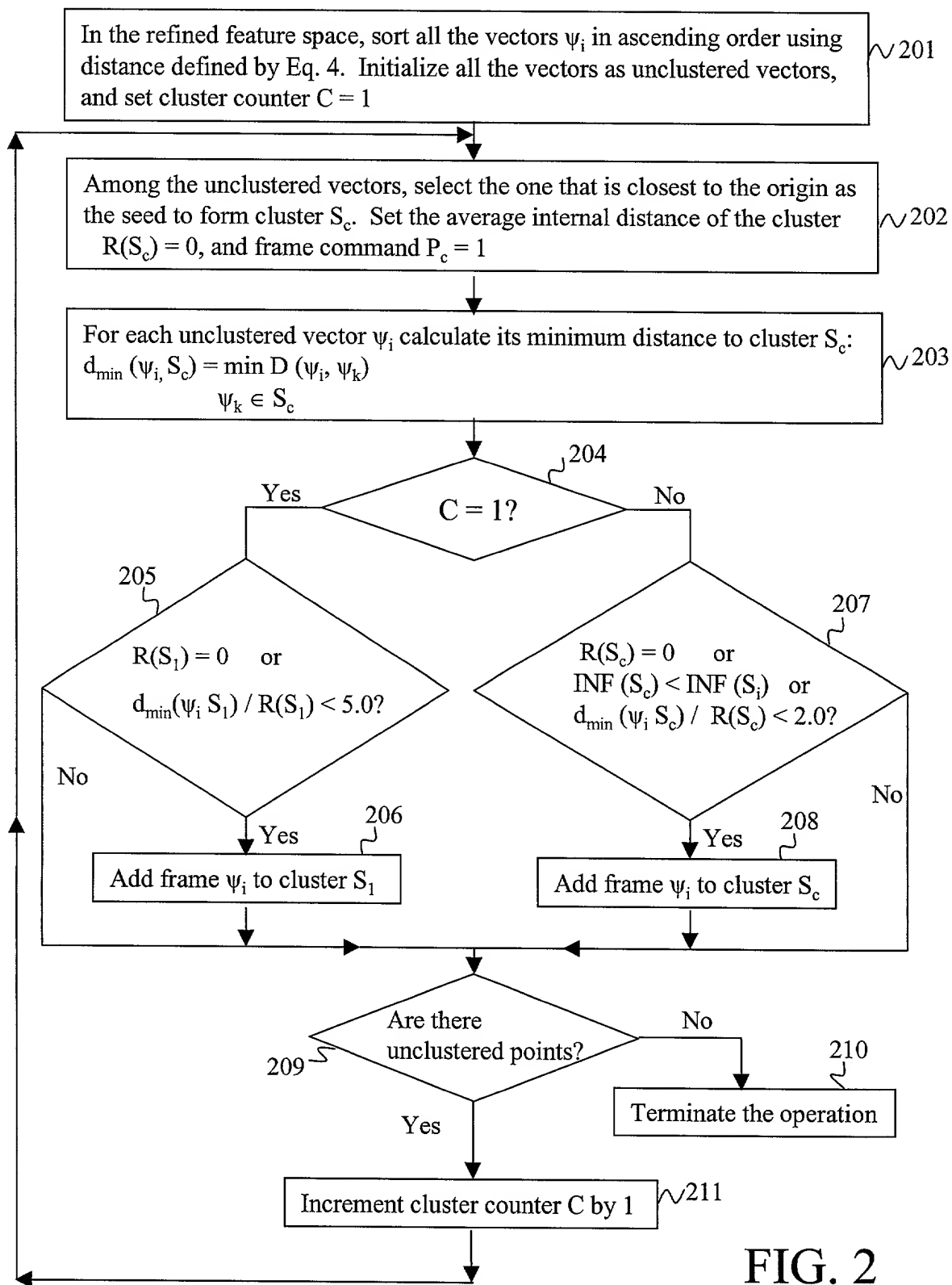


FIG. 3

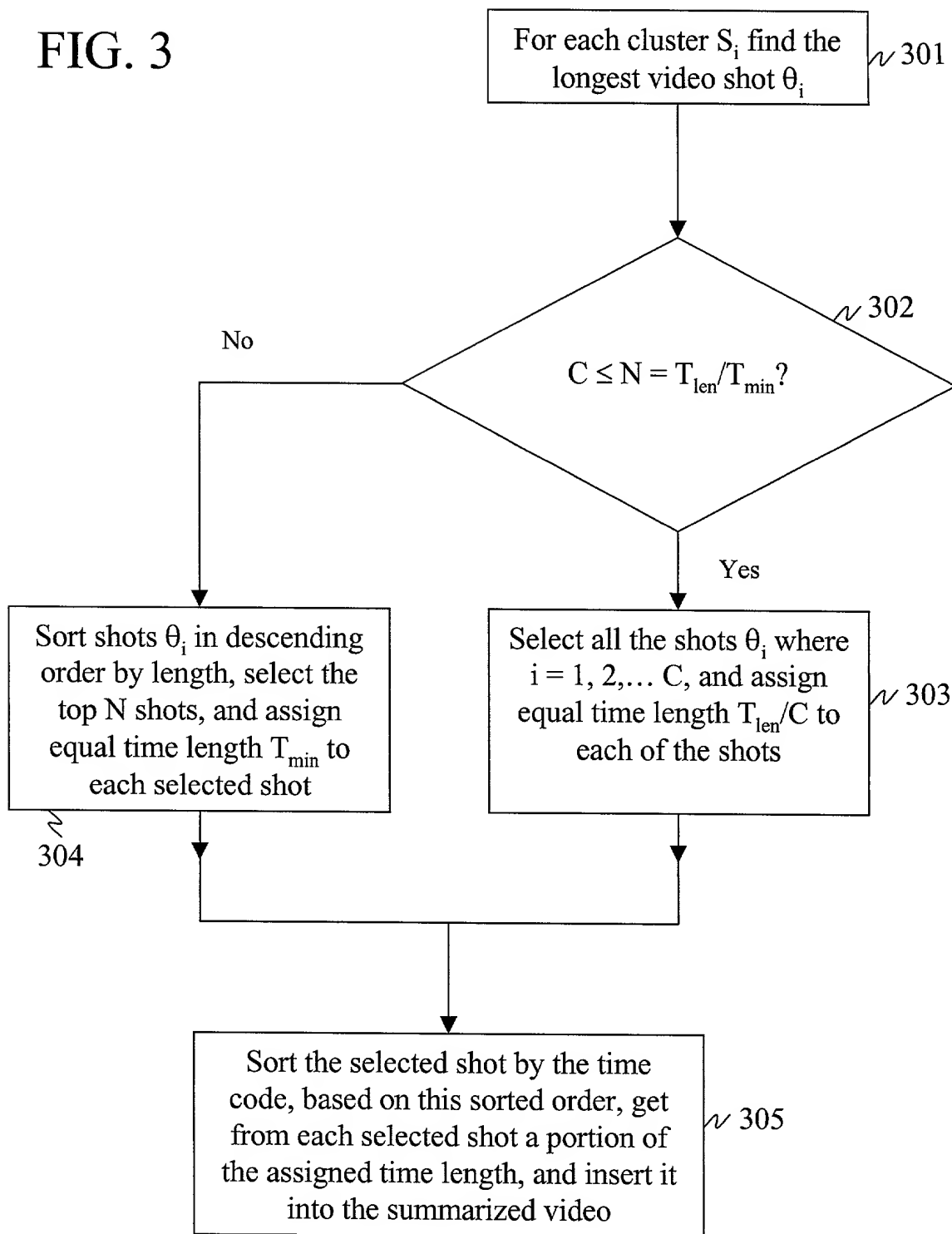


FIG. 4

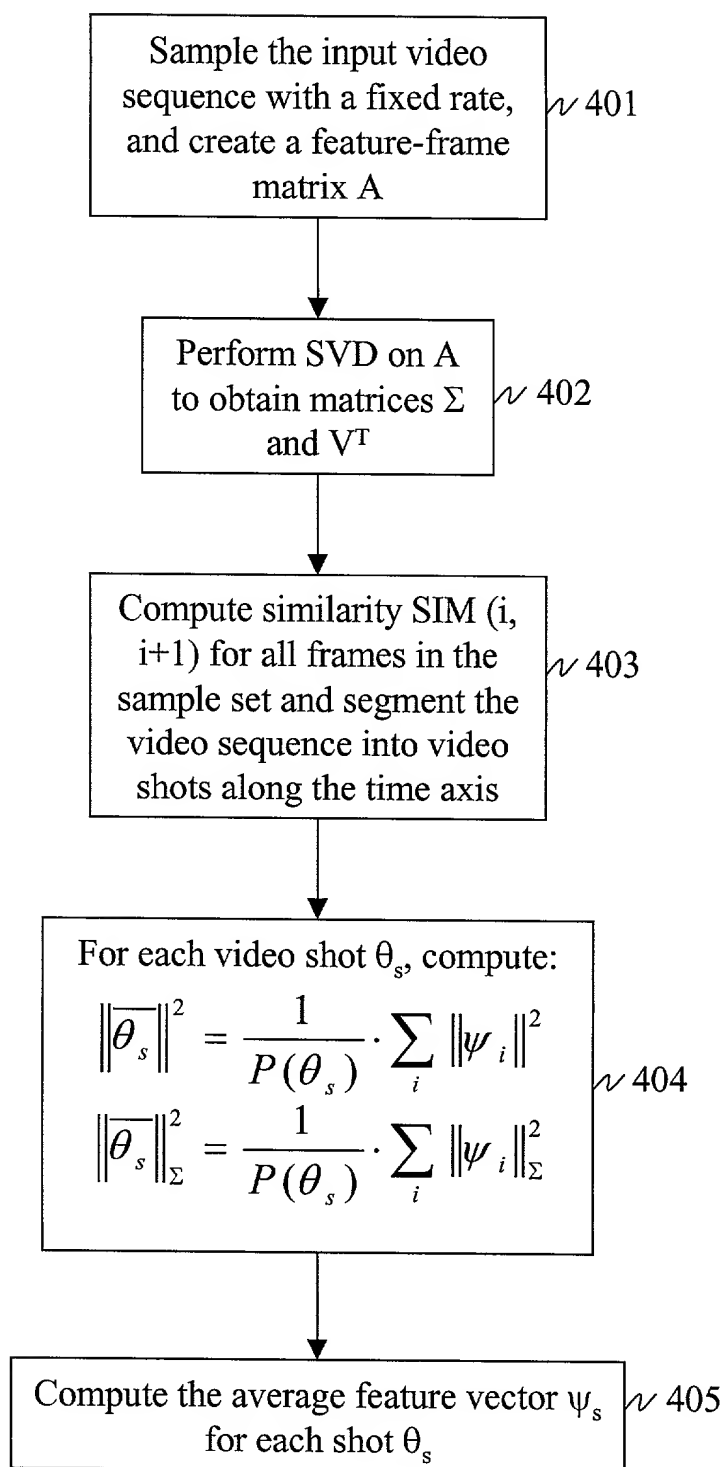
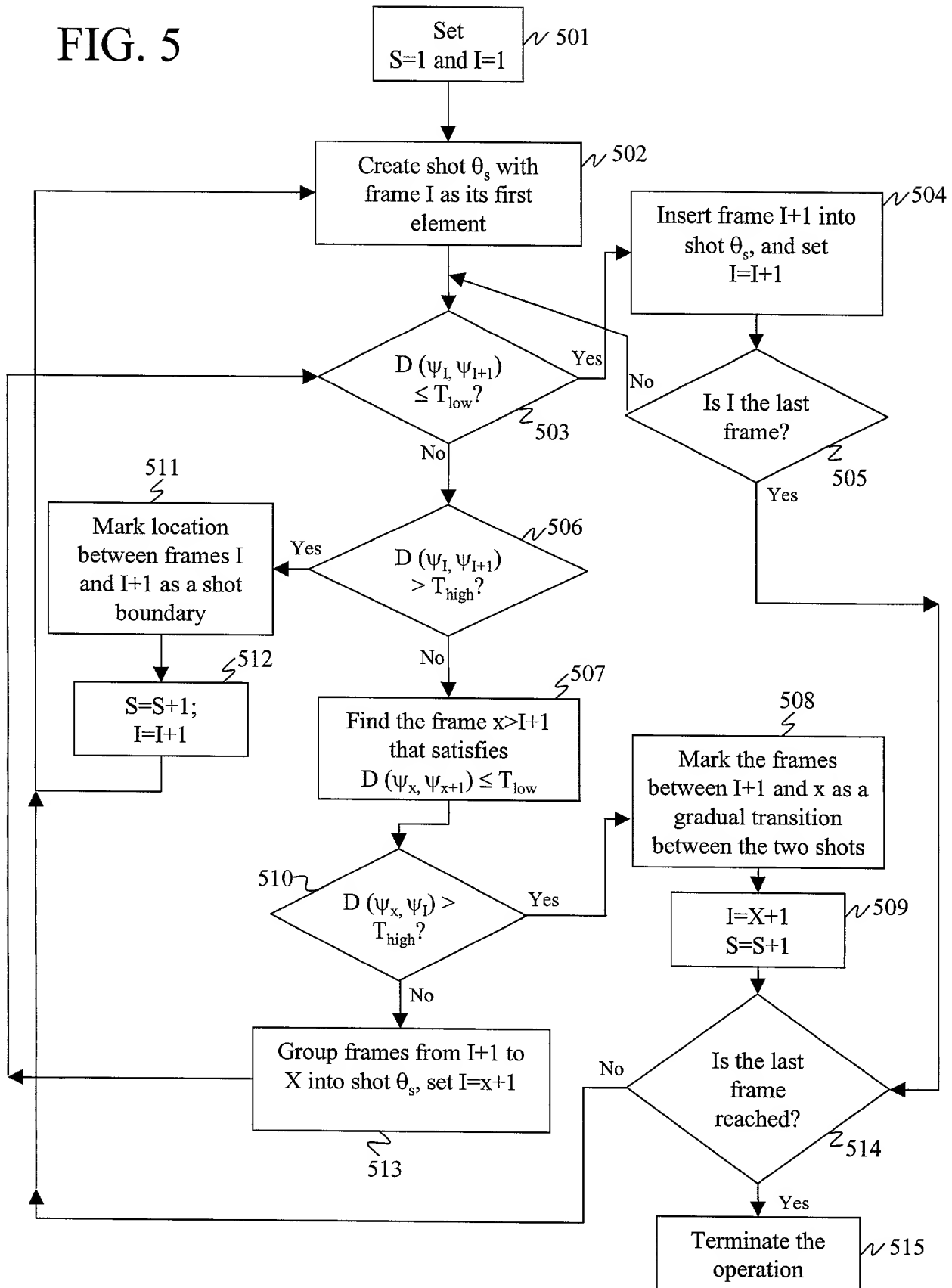




FIG. 5



**DECLARATION AND POWER OF ATTORNEY**

As a below named inventor, I hereby declare that my residence, post office address and citizenship are as stated below next to my name: that I verily believe I am the original, first and sole inventor (if only one name is listed below) or an original, first and joint inventor (if plural names are listed below) of the subject matter claimed and for which a patent is sought in the application entitled:

**METHOD AND SYSTEM FOR SEGMENTATION, CLASSIFICATION, SUMMARIZATION OF VIDEO IMAGES**

which application is:

☒ the attached application  
(for original application)

\_\_\_ application Serial No. \_\_\_\_\_  
filed \_\_\_\_\_, and amended on \_\_\_\_\_

\_\_\_\_\_  
(for declaration not accompanying application)

that I have reviewed and understand the contents of the specification of the above-identified application, including the claims, as amended by any amendment referred to above; that I acknowledge my duty to disclose information of which I am aware which is material to the examination of this application under 37 C.F.R. 1.56, that I hereby claim foreign priority benefits under Title 35, United States Code §119, §172 or §365 of any foreign application(s) for patent or inventor's certificate listed below and have also identified on said list any foreign application for patent or inventor's certificate on this invention having a filing date before that of the application on which priority is claimed:

Application Number	Country	Filing Date	Priority Claimed (yes or no)
--------------------	---------	-------------	---------------------------------

I hereby claim the benefit of Title 35, United States Code §120 of any United States application(s) listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in a listed prior United States application in the manner provided by the first paragraph of Title 35, United States Code, §112, I acknowledge my duty to disclose any material information under 37 C.F.R. 1.56 which occurred between the filing date of the prior application and the national or PCT international filing date of this application:

Application Serial No.	Filing Date	Status (patented, pending, abandoned)
60/167,230	11/24/1999	pending
60/172,379	12/17/1999	pending

I hereby appoint John H. Mion, Reg. No. 18,879; Thomas J. Macpeak, Reg. No. 19,292; Robert J. Seas, Jr., Reg. No. 21,092; Darryl Mexic, Reg. No. 23,063; Robert V. Sloan, Reg. No. 22,775; Peter D. Olexy, Reg. No. 24,513; J. Frank Osha, Reg. No. 24,625; Waddell A. Biggart, Reg. No. 24,861; Louis Gubinsky, Reg. No. 24,835; Neil B. Siegel, Reg. No. 25,200; David J. Cushing, Reg. No. 28,703; John R. Inge, Reg. No. 26,916; Joseph J. Ruch, Jr., Reg. No. 26,577; Sheldon I. Landsman, Reg. No. 25,430; Richard C. Turner, Reg. No. 29,710; Howard L. Bernstein, Reg. No. 25,665; Alan J. Kasper, Reg. No. 25,426; Kenneth J. Burchfiel, Reg. No. 31,333; Gordon Kit, Reg. No. 30,764; Susan J. Mack, Reg. No. 30,951; Frank L. Bernstein, Reg. No. 31,484; Mark Boland, Reg. No. 32,197; William H. Mandir, Reg. No. 32,156; Brian W. Hannon, Reg. No. 32,778; Abraham J. Rosner, Reg. No. 33,276; Bruce E. Kramer, Reg. No. 33,725; Paul F. Neils, Reg. No. 33,102; Brett S. Sylvester, Reg. No. 32,765; Robert M. Masters, Reg. No. 35,603; George F. Lehnigk, Reg. No. 36,359; John T. Callahan, Reg. No. 32,607 and Steven M. Gruskin, Reg. No. 36,818, my attorneys to prosecute this application and to transact all business in the Patent and Trademark Office connected therewith, and request that all correspondence about the application be addressed to **SUGHRUE, MION, ZINN, MACPEAK & SEAS, PLLC**, 1010 El Camino Real, Suite 360, Menlo Park, CA 94025.

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issuing thereon.

Date 4/20/2000

First Inventor Yihong GONG  
First Name Middle Initial Last Name

Signature [Signature]

Residence 503 Alberta Avenue, Sunnyvale, CA 94087

Post Office Address Same as above Citizenship China

Date X 04/20/2000

Second Inventor Xin  
First Name

Middle Initial

LIU  
Last Name

Signature X [Signature]

Residence 340 Auburn Way, #24, San Jose, CA 95129

Post Office Address Same as above

Citizenship China

Date \_\_\_\_\_

Third Inventor \_\_\_\_\_

First Name

Middle Initial

Last Name

Signature \_\_\_\_\_

Residence \_\_\_\_\_

Post Office Address \_\_\_\_\_

Citizenship \_\_\_\_\_

Date \_\_\_\_\_

Fourth Inventor \_\_\_\_\_

First Name

Middle Initial

Last Name

Signature \_\_\_\_\_

Residence \_\_\_\_\_

Post Office Address \_\_\_\_\_

Citizenship \_\_\_\_\_

Date \_\_\_\_\_

Fifth Inventor \_\_\_\_\_

First Name

Middle Initial

Last Name

Signature \_\_\_\_\_

Residence \_\_\_\_\_

Post Office Address \_\_\_\_\_

Citizenship \_\_\_\_\_

Date \_\_\_\_\_

Sixth Inventor \_\_\_\_\_

First Name

Middle Initial

Last Name

Signature \_\_\_\_\_

Residence \_\_\_\_\_

Post Office Address \_\_\_\_\_

Citizenship \_\_\_\_\_

Date \_\_\_\_\_

Seventh Inventor \_\_\_\_\_

First Name

Middle Initial

Last Name

Signature \_\_\_\_\_

Residence \_\_\_\_\_

Post Office Address \_\_\_\_\_

Citizenship \_\_\_\_\_